1-18. (cancelled)

19. (new) A method operative in a content delivery network (CDN) including a set of CDN servers, each CDN server provisioned with a manager process together with an application server on which one or more web applications are capable of being loaded and executed, comprising:

for each CDN server and its respective manager process, identifying values for (i) a flit-capacity, and (ii) a memory capacity, where a flit is an arbitrary unit of work representing resource usage on the CDN server;

using the values to generate a weighted mapping of web applications to manager processes for the set of CDN servers such that the flit and memory capacities for each CDN server are not exceeded; and

servicing requests at the CDN servers in proportion to the weighted mapping.

20. (new) The method as described in claim 19 wherein the weighted mapping of web applications to manager processes is also a function of application server memory capacity on each CDN server.

21. (new) The method as described in claim 19 wherein the weighted mapping of web applications to manager processes balances flits across the set of CDN servers.

22. (new) The method as described in claim 21 further including the step of re-generating the weighted mapping of web applications to manager processes for the set of CDN servers if the flit values across the set of CDN servers becomes unbalanced.

23. (new) The method as described in claim 19 wherein the flit represents non-bandwidth resource usage at a CDN server.

24. (new) The method as described in claim 23 wherein the flit is CPU utilization.

25. (new) The method as described in claim 19 wherein the weighted mapping of web applications to manager processes requires a given web application to be loaded onto the CDN server.

26. (new) The method as described in claim 19 wherein the weighted mapping of web applications to manager processes requires a given web application to be unloaded from the CDN server.

27. (new) The method as described in claim 19 wherein the set of CDN servers are co-located.

28. (new) The method as described in claim 19 wherein a given request is serviced at by an instance of a web application loaded and executing on a given CDN server.